Supplemental data

GimmeMotifs: a *de novo* motif prediction pipeline for ChIP-sequencing experiments

Simon J. van Heeringen[1,*], Gert Jan C. Veenstra[1]

1 Radboud University Nijmegen, Department of Molecular Biology, Faculty of Science, Nijmegen Centre for Molecular Life Sciences, 6500 HB Nijmegen, The Netherlands

∗ E-mail: s.vanheeringen@ncmls.ru.nl

Supplemental Methods

Example output

Examples of GimmeMotifs results and output files are included for four different human transcription factor ChIP-seq datasets: NRSF, STAT1, CTCF (Jothi *et al.*, 2008) and p63 Kouwenhoven *et al.* (2010). GimmeMotifs was run with the 'large' analysis setting and default parameters. For the datasets from Jothi *et al.* the same amount of sequences were used as in their analyses, to enable comparison of the results.

Motif programs and parameters

Currently BioProspector (Liu *et al.*, 2001), GADEM (Li, 2009), Improbizer (Ao *et al.*, 2004), MDmodule (Liu *et al.*, 2002), MEME (Bailey *et al.*, 2009), MoAn (Valen *et al.*, 2009), MotifSampler (Thijs *et al.*, 2001), trawler (Ettwiller *et al.*, 2007) and Weeder (Pavesi *et al.*, 2004) are supported.
Except for the width, which is varied according to the GimmeMotifs `analysis` parameter, mostly default settings are used for each program. Single- or double-stranded analysis is specified, if possible, according to GimmeMotifs settings. If a background file is required or recommended then a background file is generated based on the input sequences using the matched_genomic method (see below). This background file is used for Improbizer, MoAn and BioProspector.

Matched genomic background

For all peaks/sequences the closest transcription start site (TSS) of a gene is determined. For each sequence in the input, multiple sequences are selected randomly from the genome (default 10), on the same chromosome if possible, with a similar distance to the TSS of a gene. In this manner the distribution of the background sequences with respect to gene TSS's will be similar to the input sequences. This background set is optional for less well-annotated genomes, as it is dependent on an accurate gene annotation.

Clustering

All significant motifs are clustered using an iterative procedure. Pairwise comparisons are performed for all motifs using the WIC score. The two most similar motifs are merged, and an average motif is computed. The average column frequencies are based on the weighted column frequencies of the two individual PFMs. Therefore, a motif with a large amount of occurences has a larger influence on the new averaged motif. The pairwise comparison scores of the new average motif to all other motifs are calculated, and the two most similar motifs are again merged. This procedure is repeated iteratively until the best scoring alignment does not reach a predefined threshold. By default this threshold is a WIC p-value of ≤ 0.05. This is an empirical p-value which is calculated based on the maximum WIC score and the length of the motif using the method of Sandelin and Wasserman, based on simulated PFMs (Sandelin and Wasserman, 2004). 10,000 random PFMs were generated using the JASPAR website (`http://jaspar.cgb.ki.se/`).

Statistics

The statistics (enrichment, hypergeometric p-values, ROC and MNCP) are calculated by comparing the validation set (sequences not used for motif prediction) and a background set of sequences. GimmeMotifs uses two different backgrounds by default. One is a set of randomly generated sequences. A 1st order Hidden Markov Model is trained on the input data and used to generate sequences with similar dinucleotide frequencies. Additionally, GimmeMotifs randomly selects genomic sequences, taking into account the position of the peaks relative to the transcription start site (TSS) of genes. For example, if half of the input sequences are located in the proximal promoter, half of

the background set will correspondingly be selected from proximal promoter regions. All p-values are corrected by Benjamini-Hochberg multiple testing correction where applicable.

Additional Tools

The prediction of an accurate transcription factor binding motif is often one of the first steps in ChIP-seq data analysis. However, once a motif has been determined, this predicted motif often needs to be used to evaluate other datasets or specific subsets of the dataset. GimmeMotifs includes several command-line tools to facilitate these typical analyses. For example, the weight matrix predicted in the pipeline can be used to scan sequences with the predicted motif using an optimized motif-specific cutoff. All the steps included in the pipeline are also available as separate scripts, which can perform tasks such as conversion of a BED file with genomic coordinates to a FASTA file with sequences, ROC and MNCP evaluation, generation of background sequences and clustering of motifs.

Benchmarks

Benchmark datasets were retrieved from the publications of Chen *et al.* (2008) (mouse: Esrrb, Nanog, Oct4, Sox2, CTCF, E2f1, Smad1, Tcfcp2l1, Zfx, Klf4, c-Myc, n-Myc and STAT3) and Valouev *et al.* (2008); Jothi *et al.* (2008) (human: SRF, GABP, NRSF, CTCF and STAT1). For each dataset 3000 peaks of length 200 were randomly chosen from the total set of peaks. GimmeMotifs was run with default parameters, with the analysis set to 'xl'. With these settings 20% of the peaks (600 in total) were used for prediction, and 2400 were used for validation. Of the prediction set, only the first 300 peaks (because of performance reasons) were submitted to the web interface of W-ChipMotifs (Jin *et al.*, 2009) and SCOPE (Carlson *et al.*, 2007) with default settings. All statistics (ROC AUC, MNCP) were calculated using the 2400 sequences not used for motif prediction.

Supplemental Tables

Table S1. Benchmark datasets: ROC AUC

	GimmeMotifs			W-ChIPmotifs			SCOPE		
	Random	**Genomic**	**Mean**	**Random**	**Genomic**	**Mean**	**Random**	**Genomic**	**Mean**
Mouse									
CTCF	0.964	0.963	0.964	0.963	0.961	0.962	0.516	0.556	0.536
E2f1	0.646	0.774	0.710	0.666	0.759	0.713	0.561	0.731	0.646
Esrrb	0.911	0.922	0.917	0.912	0.923	0.917	0.559	0.691	0.625
Klf4	0.890	0.903	0.896	0.890	0.893	0.892	0.723	0.786	0.754
Nanog	0.750	0.702	0.726	0.750	0.687	0.719	0.600	0.579	0.590
Oct4	0.740	0.692	0.716	0.736	0.686	0.711	0.594	0.601	0.598
STAT3	0.762	0.774	0.768	0.759	0.775	0.767	0.596	0.702	0.649
Smad1	0.723	0.681	0.702	0.724	0.677	0.701	0.592	0.588	0.590
Sox2	0.867	0.838	0.852	0.868	0.829	0.848	0.628	0.605	0.616
Tcfcp2l1	0.892	0.924	0.908	0.885	0.906	0.896	0.558	0.524	0.541
Zfx	0.711	0.892	0.802	0.666	0.885	0.775	0.605	0.661	0.633
c-Myc	0.767	0.875	0.821	0.740	0.885	0.813	0.632	0.741	0.687
n-Myc	0.728	0.860	0.794	0.672	0.861	0.766	0.643	0.734	0.688
Human									
CTCF	0.926	0.942	0.934	0.926	0.943	0.934	0.514	0.561	0.537
GABP	0.918	0.923	0.920	0.915	0.913	0.914	0.649	0.629	0.639
NRSF	0.906	0.936	0.921	0.903	0.922	0.913	0.540	0.534	0.537
SRF	0.701	0.720	0.711	0.700	0.724	0.712	0.568	0.679	0.624
STAT1	0.882	0.888	0.885	0.880	0.887	0.884	0.516	0.559	0.538
Median	**0.816**	**0.845**	**0.830**	**0.809**	**0.840**	**0.824**	**0.588**	**0.637**	**0.613**

Shown is the ROC AUC of the best performing motif for GimmeMotifs, W-ChipMotifs and SCOPE for each benchmark dataset. The ROC AUC is calculated for two different background datasets (random and matched genomic) and the mean ROC AUC is calculated across the two background sets.

Table S2. Benchmark datasets: MNCP

	GimmeMotifs			W-ChIPmotifs			SCOPE		
	Random	**Genomic**	**Mean**	**Random**	**Genomic**	**Mean**	**Random**	**Genomic**	**Mean**
Mouse									
CTCF	9.575	8.385	8.980	9.584	8.327	8.955	1.141	1.076	1.109
E2f1	2.278	2.527	2.403	2.408	2.616	2.512	2.215	1.481	1.848
Esrrb	7.439	7.087	7.263	7.445	7.097	7.271	1.861	1.458	1.659
Klf4	6.056	4.879	5.468	6.006	4.838	5.422	3.201	3.153	3.177
Nanog	4.311	3.396	3.854	3.773	2.739	3.256	1.259	1.569	1.414
Oct4	6.243	5.133	5.688	6.239	5.124	5.682	1.400	1.789	1.594
STAT3	4.157	4.271	4.214	3.944	3.980	3.962	1.889	2.162	2.025
Smad1	4.104	3.149	3.627	3.933	2.877	3.405	1.413	1.533	1.473
Sox2	5.891	5.005	5.448	5.888	4.979	5.433	1.622	1.830	1.726
Tcfcp2l1	6.021	7.045	6.533	5.796	6.310	6.053	1.080	1.359	1.220
Zfx	2.340	4.787	3.564	2.492	4.549	3.520	1.480	1.531	1.506
c-Myc	3.743	5.112	4.427	3.208	5.431	4.319	2.249	1.744	1.997
n-Myc	3.078	4.400	3.739	2.706	4.131	3.418	2.129	1.803	1.966
Human									
CTCF	8.742	8.727	8.735	8.715	8.700	8.707	1.193	1.087	1.140
GABP	6.744	5.439	6.091	6.679	5.246	5.962	1.396	1.763	1.580
NRSF	8.870	9.003	8.937	8.874	9.012	8.943	1.077	1.069	1.073
SRF	4.707	4.553	4.630	4.673	4.549	4.611	1.730	1.495	1.612
STAT1	7.351	7.150	7.251	7.357	7.152	7.254	1.180	1.085	1.132
Median	**5.647**	**5.558**	**5.603**	**5.540**	**5.425**	**5.483**	**1.640**	**1.610**	**1.625**

Shown is the MNCP of the best performing motif for GimmeMotifs, W-ChipMotifs and SCOPE for each benchmark dataset. The MNCP is calculated for two different background datasets (random and matched genomic) and the mean MNCP is calculated across the two background sets.

Table S3. Running time of different algorithms

Algorithm	Running time to predict motifs (h:mm)		
	NRSF (5,813)	**CTCF (26,814)**	**CTCF (top 500)**
BioProspector	0:04	0:05	0:01
GADEM	1:03	0:45	0:05
Improbizer	0:12	0:12	0:02
MDmodule	0:01	0:01	0:01
MEME	5:06	5:56	0:04
MoAn	87:11	90:41	9:50
MotifSampler	1:10	1:02	0:04
trawler	0:02	0:08	0:01
Weeder	1:05	1:08	0:08
GimmeMotifs	0:44	2:32	0:05

This table gives an *indication* of the running time of GimmeMotifs and the individual motif prediction algorithms. Two datasets from Jothi *et al.* were used as input for GimmeMotifs: NRSF and CTCF, with the amount of sequences shown in brackets. In addition we also used a selection of the highest 500 peaks of the CTCF dataset. GimmeMotifs was run with default settings: analysis size 'medium' and maximum 1000 sequences used for motif prediction. The running time of each individual algorithm is shown, as well as the time that GimmeMotifs takes for retrieving sequences, clustering motifs, determining significance etc. The GimmeMotifs benchmark time does not include the motif prediction, and will vary with the amount of motifs predicted. The analyses were run on a 12-core 2100Mhz AMD Opteron machine with 64Gb internal memory. As most of the algorithms are run in parallel (except GADEM, MoAn and trawler), the time will vary with the number of CPUs available.

Supplemental Figures

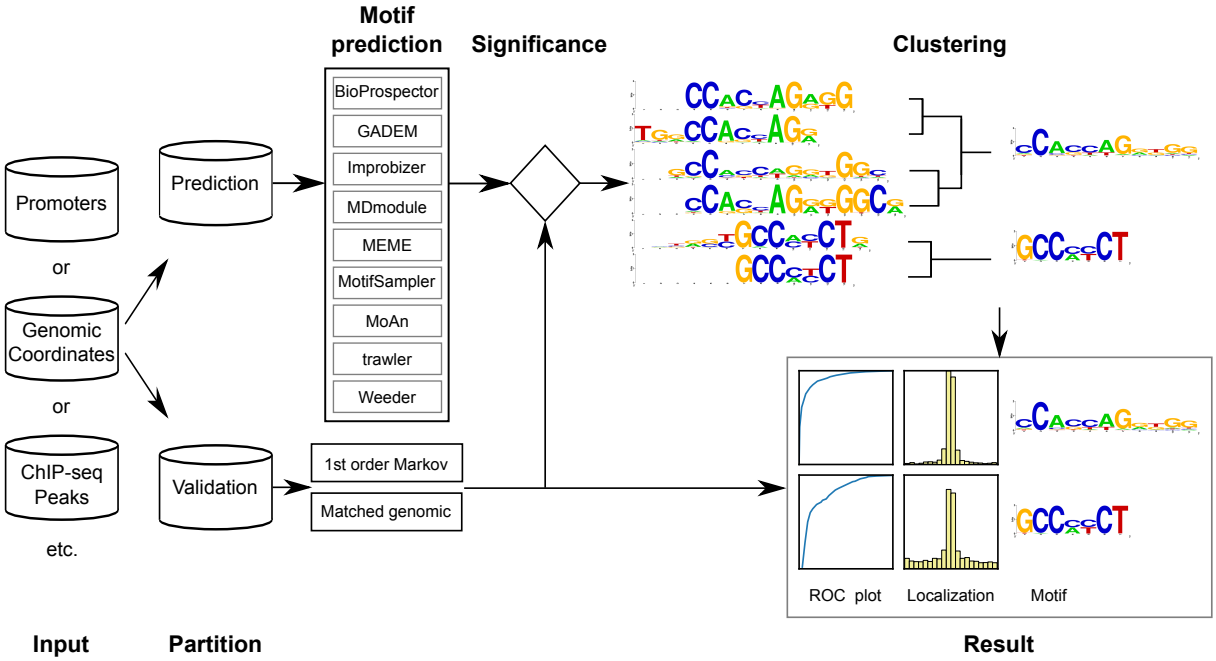

Figure S1. A flowchart of GimmeMotifs.

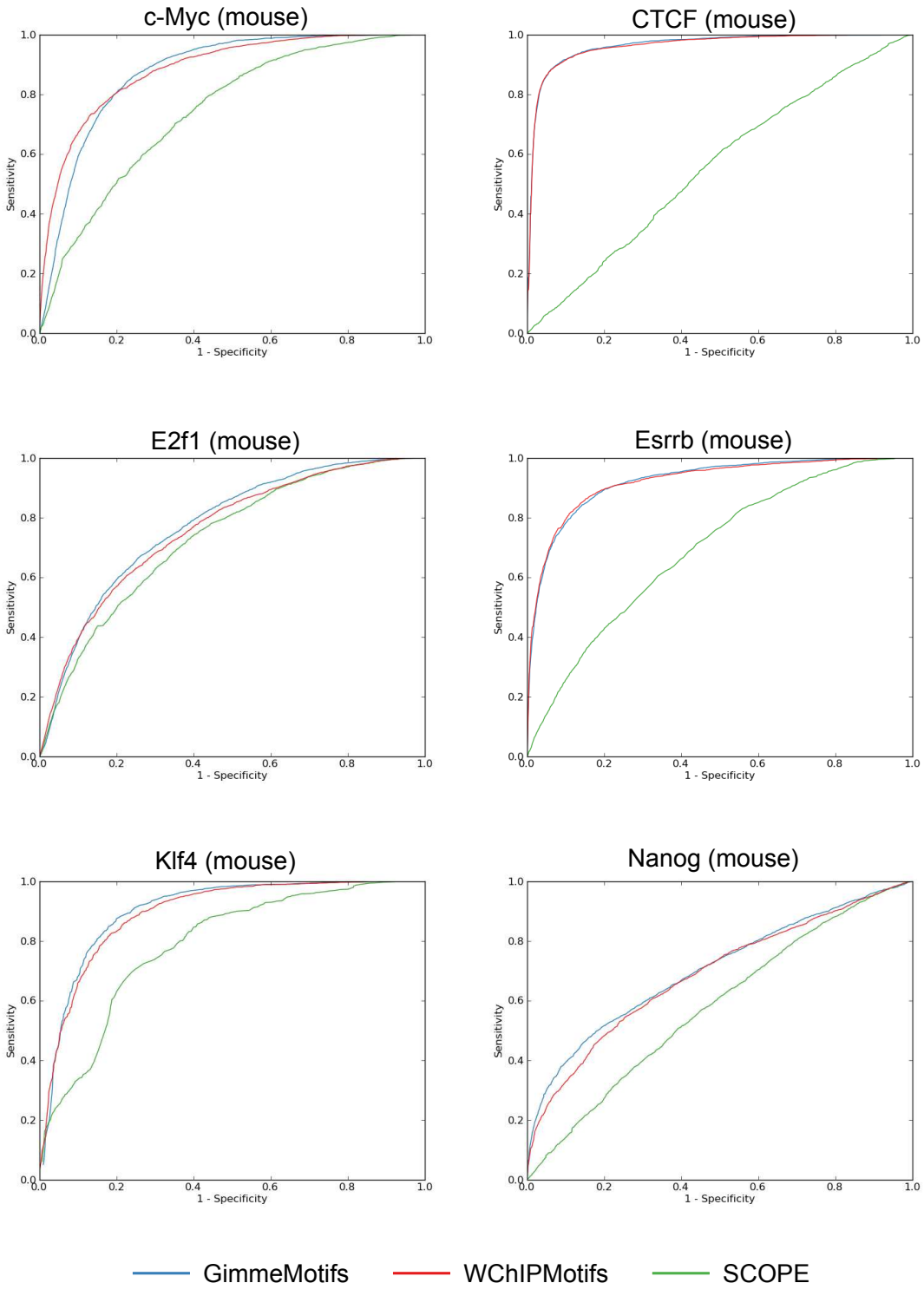

Figure S2. ROC curves for benchmark datasets shown in Supplementary Table 1

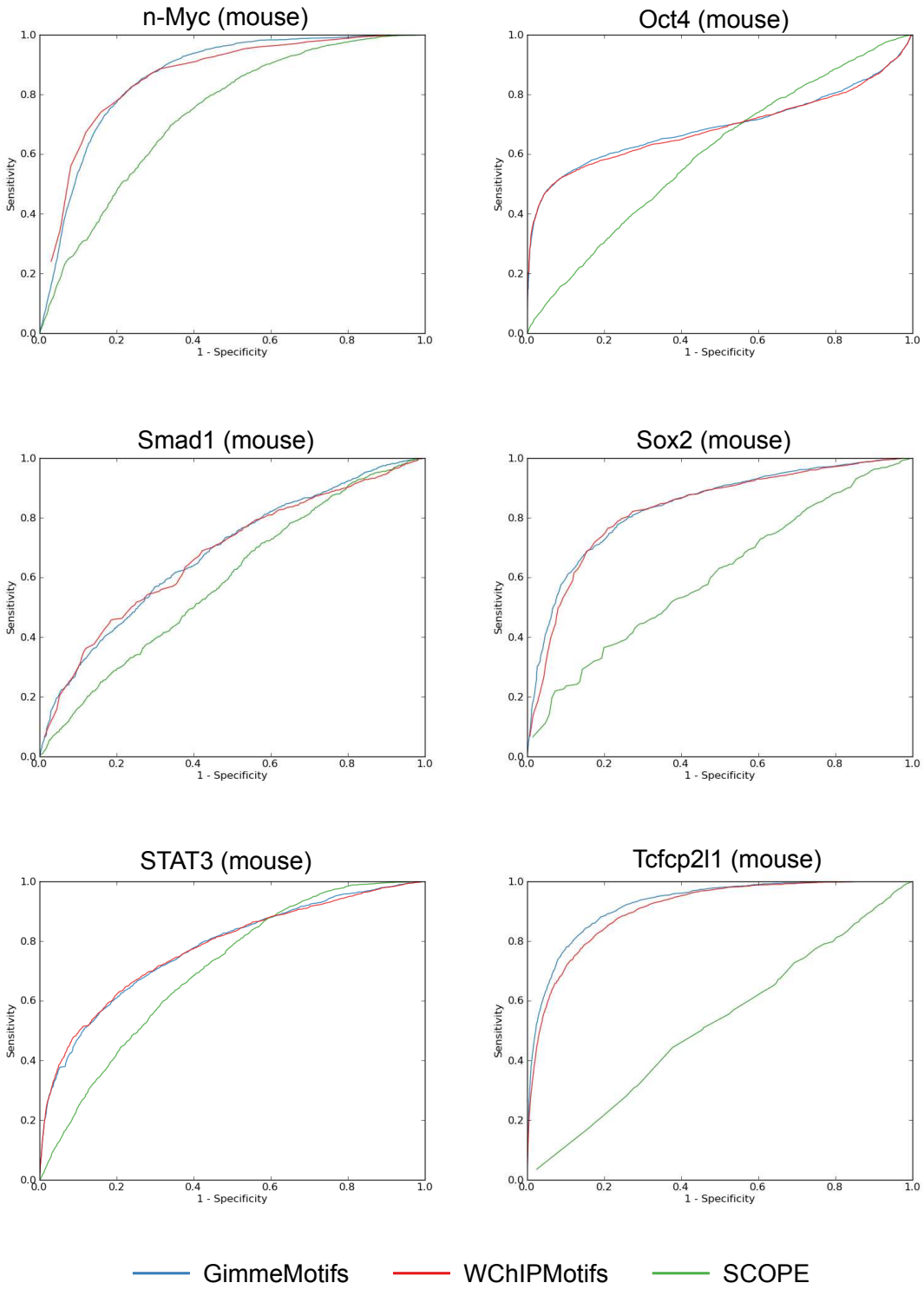

Figure S3. ROC curves for benchmark datasets shown in Supplementary Table 1

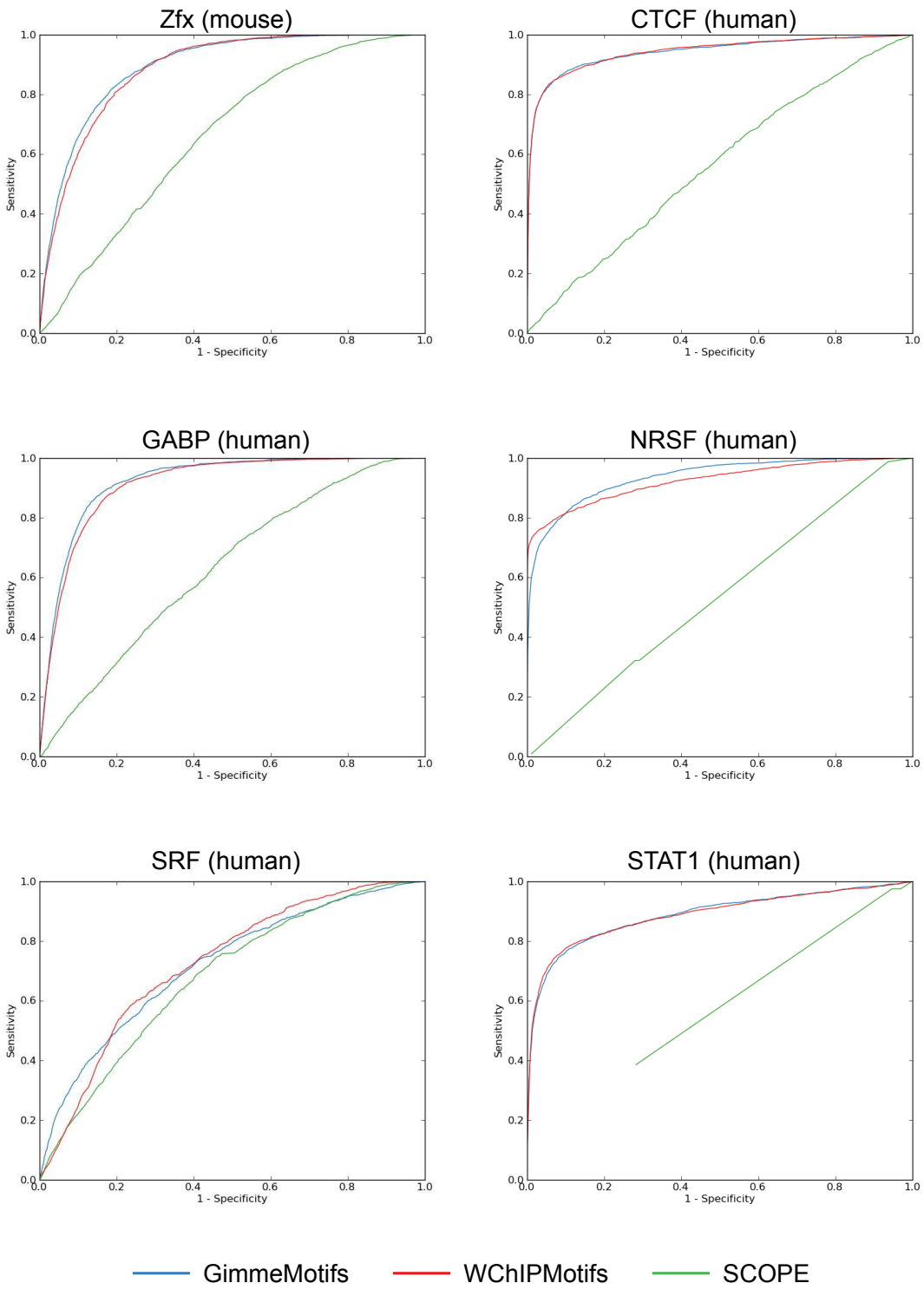

Figure S4. ROC curves for benchmark datasets shown in Supplementary Table 1

References

Ao, W., Gaudet, J., Kent, W. J., Muttumu, S., and Mango, S. E. (2004). Environmentally induced foregut remodeling by PHA-4/FoxA and DAF-12/NHR. *Science*, **305**(5691), 1743–1746.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucl. Acids Res.*, **37**(suppl_2), W202–208.

Carlson, J. M., Chakravarty, A., DeZiel, C. E., and Gross, R. H. (2007). SCOPE: a web server for practical de novo motif discovery. *Nucl. Acids Res.*, **35**(suppl_2), W259–264.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V. B., Wong, E., Orlov, Y. L., Zhang, W., and Jiang, J. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell*, **133**(6), 1106–1117.

Ettwiller, L., Paten, B., Ramialison, M., Birney, E., and Wittbrodt, J. (2007). Trawler: de novo regulatory motif discovery pipeline for chromatin immunoprecipitation. *Nat Meth*, **4**(7), 563–565.

Jin, V. X., Apostolos, J., Nagisetty, N. S. V. R., and Farnham, P. J. (2009). W-ChIPMotifs: a web application tool for de novo motif discovery from ChIP-based high-throughput data. *Bioinformatics*, **25**(23), 3191–3193.

Jothi, R., Cuddapah, S., Barski, A., Cui, K., and Zhao, K. (2008). Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucl. Acids Res.*, **36**(16), 5221–5231.

Kouwenhoven, E. N., van Heeringen, S. J., Tena, J. J., Oti, M., Dutilh, B. E., Alonso, M. E., de la Calle-Mustienes, E., Smeenk, L., Rinne, T., Parsaulian, L., Bolat, E., Jurgelenaite, R., Huynen, M. A., Hoischen, A., Veltman, J. A., Brunner, H. G., Roscioli, T., Oates, E., Wilson, M., Manzanares, M., Gmez-Skarmeta, J. L., Stunnenberg, H. G., Lohrum, M., van Bokhoven, H., and Zhou, H. (2010). Genome-Wide profiling of p63 DNA-Binding sites identifies an element that regulates gene expression during limb development in the 7q21 SHFM1 locus. *PLoS Genet*, **6**(8), e1001065.

Li, L. (2009). GADEM: a genetic algorithm guided formation of spaced dyads coupled with an EM algorithm for motif discovery. *Journal of computational biology : a journal of computational molecular cell biology*, **16**(2), 317–329. PMID: 19193149 PMCID: 2756050.

Liu, X., Brutlag, D. L., and Liu, J. S. (2001). BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 127–138. PMID: 11262934.

Liu, X. S., Brutlag, D. L., and Liu, J. S. (2002). An algorithm for finding protein-DNA binding sites with applications to chromatin-immunoprecipitation microarray experiments. *Nat Biotech*, **20**(8), 835–839.

Pavesi, G., Mereghetti, P., Mauri, G., and Pesole, G. (2004). Weeder web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucl. Acids Res.*, **32**(suppl_2), W199–203.

Sandelin, A. and Wasserman, W. W. (2004). Constrained binding site diversity within families of transcription factors enhances pattern discovery bioinformatics. *Journal of Molecular Biology*, **338**(2), 207–215. PMID: 15066426.

Thijs, G., Lescot, M., Marchal, K., Rombauts, S., Moor, B. D., Rouz, P., and Moreau, Y. (2001). A higher-order background model improves the detection of promoter regulatory elements by gibbs sampling. *Bioinformatics (Oxford, England)*, **17**(12), 1113–1122. PMID: 11751219.

Valen, E., Sandelin, A., Winther, O., and Krogh, A. (2009). Discovery of regulatory elements is improved by a discriminatory approach. *PLoS Comput Biol*, **5**(11), e1000562.

Valouev, A., Johnson, D. S., Sundquist, A., Medina, C., Anton, E., Batzoglou, S., Myers, R. M., and Sidow, A. (2008). Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat Meth*, **5**(9), 829–834.